A survey of social web mining applications for disease outbreak detection

Gema Bello-Orgaz, Julio Hernandez-Castro and David Camacho

Abstract Social Web Media is one of the most important sources of big data to extract and acquire new knowledge. Social Networks have become an important environment where users provide information of their preferences and relationships. This information can be used to measure the influence of ideas and the society opinions in real time, being very useful on several fields and research areas such as marketing campaigns, financial prediction or public healthcare among others. Recently, the research on artificial intelligence techniques applied to develop technologies allowing monitoring web data sources for detecting public health events has emerged as a new relevant discipline called Epidemic Intelligence. Epidemic Intelligence Systems are nowadays widely used by public health organizations like monitoring mechanisms for early detection of disease outbreaks to reduce the impact of epidemics. This paper presents a survey on current data mining applications and web systems based on web data for public healthcare over the last years. It tries to take special attention to machine learning and data mining techniques and how they have been applied to these web data to extract collective knowledge from Twitter

Gema Bello-Orgaz
Escuela Politecnica Superior, Universidad Autonoma de Madrid, Francisco Tomas y Valiente 11, 28049 Madrid, Spain, e-mail: gema.bello@uam.es

Julio Hernandez-Castro
School of Computing, University of Kent, Cornwallis South Building, Canterbury CT2 7NF, UK
e-mail: J.C.Hernandez-Castro@kent.ac.uk

David Camacho
Escuela Politecnica Superior, Universidad Autonoma de Madrid, Francisco Tomas y Valiente 11, 28049 Madrid, Spain, e-mail: david.camacho@uam.es

1 Introduction

Web is one of the most important sources of data in the world producing amounts of public information. The exponentially increasing of websites and online web services in the last years has allowed new interdisciplinary challenges for several fields and computer science, such as marketing campaigns [8] [3], financial prediction [2] or public healthcare [10] [17] [7], among others. Recently, the research on artificial intelligence techniques applied to develop technologies allowing monitoring web data sources for detecting public health events has been emerged as a new relevant discipline called Epidemic Intelligence (EI).

EI can be defined as the early identification, assessment and verification of potential public health risks [25], and the timely dissemination of the appropriate alerts. This discipline includes surveillance techniques such as automated and continuous analysis of unstructured free text information available on Web from social networks, blogs, digital news media or official sources. Surveillance systems are nowadays widely used by public health organizations such as World Health Organization (WHO) or the European Centre for Disease Prevention and Control (ECDC) [16]. Tracking and monitoring mechanisms for early detection are critical in reducing the impact of epidemics giving a rapid response. For instance, several of these systems can be able to discover early events of the disease breakout during the A(H1N1) influenza pandemic in 2009 [11].

Traditional epidemic surveillance systems are implemented from virology and clinical data, which is manually collected, and often these traditional systems have a delay reporting the emerging diseases. But in situations like epidemic outbreaks, real-time feedback and a rapid response is critical. Social Web media is a profitable medium to extract the society opinion in real time. Blogs, micro-blogs (Twitter), and social networks (Facebook) enable people to publish their personal opinions in real time, including geo-information about their current locations. These big data with situation and context aware information about the users provide a useful source for public healthcare. However, the extraction of information from web is a difficult task due to its unstructured definition, high heterogeneity, and dynamically changing nature. Because of this diversity in the data format, several computational methods are required for its processing and analysing [17] (data mining, natural language processes (NLP), knowledge extraction, context awareness, etc...).

This paper presents a survey on current data mining applications and web systems based on web data for public healthcare over the last years. It tries to take special attention to machine learning and data mining techniques, and how they have been applied to these web data to extract collective knowledge from social networks like Twitter. The rest of the paper has been structured as follows: Section 2 shows the state of the art of the existing Epidemic Intelligence Systems. Section 3 describes the different web mining techniques used to detect disease outbreaks. Section 4 provides an overview of Twitter applications for monitoring and predicting epidemic and their experimental results. Finally, the last section presents a discussion of the main features extracted from this survey.

2 Epidemic Intelligence Systems for public healthcare

Nowadays, large amounts of emergency and health data are increasingly coming from a large range of web and social media sources. This information can be very useful for disease surveillance and early outbreak detection, and several public web surveillance projects in this field have emerged over the recent years.

One of the earliest surveillance systems is the Global Public Health Intelligence Network (GPHIN) [24] developed by Public Health Agency of Canada in collaboration with WHO. It is a secure web-based multilingual warning tool that is continuously monitoring and analysing global media data sources to identify information about disease outbreaks and other events related to public healthcare. The information is filtered for relevancy by an automated process, and categorized based on a specific taxonomy of categories. Then this information is analysed by Public Health Agency of Canada GPHIN officials. From 2002 to 2003 years, this surveillance system was able to detect the outbreak of Severe Acute Respiratory Syndrome (SARS).

Since 2006, BioCaster [11] is an operational ontology-based system for monitoring online media data. This system is based on text mining techniques for detecting and tracking the infectious disease outbreaks through the search of linguistic signals. The system continuously analyses documents reported from over 1700 RSS feeds, Google News, WHO, ProMED-mail, and the European Media Monitor, among others providers. The extracted text are classified for topical relevance and plots them onto a Google map using geo-information. The system consists of four main stages: topic classification, named entity recognition(NER), disease/location detection and event recognition. In the first stage, the texts are classified into relevant or non-relevant categories using to train a naive Bayes classifier. Then, for relevant document corpus are search entities of interest from 18 concept types based on the BioCaster ontology [12] related to diseases, viruses, bacteria, locations and symptoms, see Figure 1.

HealthMap project [5] is a global disease alert map which uses data from different sources such as Google News, expert-curated discussion such as ProMED-mail, and official organization reports such as World Health Organization (WHO) or Euro Surveillance. This is an automated real-time system that monitors, organizes, integrates, filters, visualizes, and disseminates online information about emerging diseases as can be seen in Figure 2.

Other system which collects news from the Web, related to human and animal health, and plot the data on a Google Maps mashupare is EpiSpider [18]. This tool automatically extracts infectious disease outbreak information from several sources including ProMed-mail and medical web sites, and it is used as surveillance system by public healthcare organizations, several universities and health research organization. Additonally, this system automatically converted the topic and location information of the reports into RSS feeds.

Other public health surveillance system used by a Public Health Organization (The European Centre of Disease Prevention and Control) is MedISys [23] monitoring human and animal infectious diseases, as well as chemical, biological, radiological and nuclear (CBRN) threats in open-source media. MedISys automatically

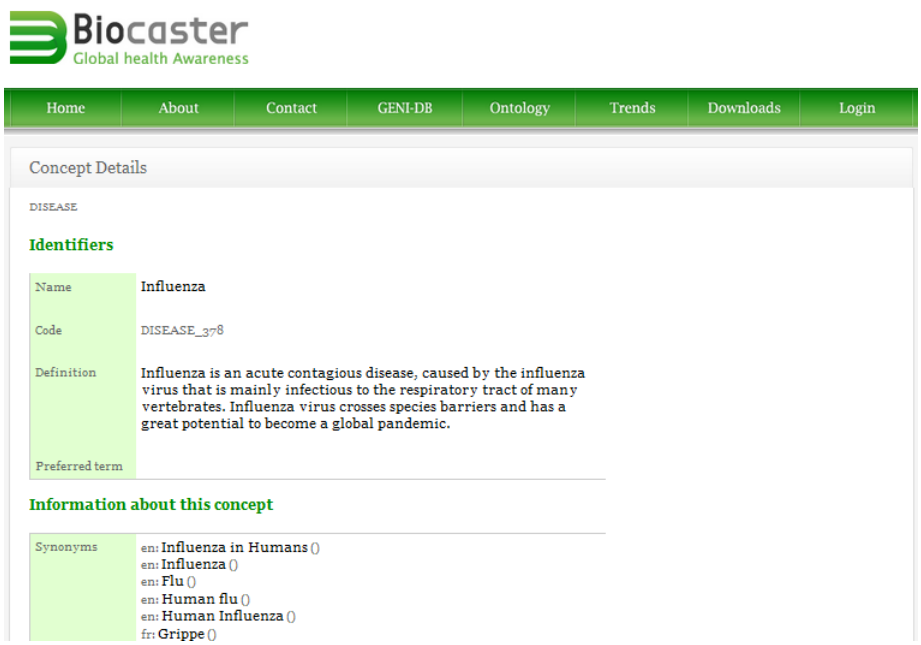

Fig. 1 BioCaster ontology related to diseases, viruses, bacteria, locations and symptoms. Screenshot taken from the BioCaster Health Monitor Web (http://born.nii.ac.jp), online accessed on 18th March 2014.

collects articles concerning public health in various languages from news, which are classified according to pre-defined categories as can be seen in Figure 3. Users can display world maps in which event locations are highlighted as well as statistics on the reporting about diseases, countries and combinations of them, also can apply filters for language, disease or location.

A specific and extensive application of predictive analytic techniques to public health approach are the monitoring systems of influenza through Web and social media. Google Flu Trends [6] uses Google search data to estimate flu activity during two weeks giving an early detection of disease activity, see Figure 4. This web service correlates search term frequency with influenza statistics reported by the Centers for Disease Control and Prevention (CDC), and it enables a quicker response in a potential pandemic of influenza, thus reducing its impact. Internet users perform search queries [15] and post entries in blogs using terms related to influenza illness as its diagnosis and symptoms. An increase or decrease in the number of illness searches and posts in blogs, reflects a higher or lower potential outbreak focus for influenza illness and can therefore be used to monitor it.

Finally all the systems mentioned together with their main characteristics are listed in Table 1.

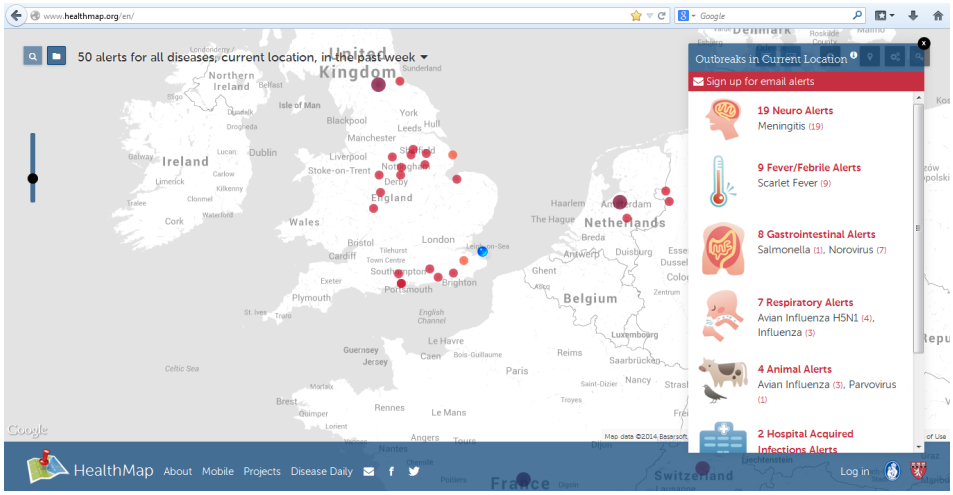

Fig. 2 HealthMap global disease alert map showing information about emerging diseases. Screenshot taken from the HealthMap Web (http://www.healthmap.org/en/), online accessed on 18th March 2014.

Fig. 3 MedISys system displaying a collect of articles concerning measles in various languages. Screenshot taken from the MedISys Web (http://medusa.jrc.it/medisys/homeedition/en/home.html), online accessed on 18th March 2014.

3 Web Mining solutions for disease outbreaks detection

The problem of detecting and tracking epidemic outbreaks through social media can be defined as the task of extracting relevant knowledge about the epidemics in

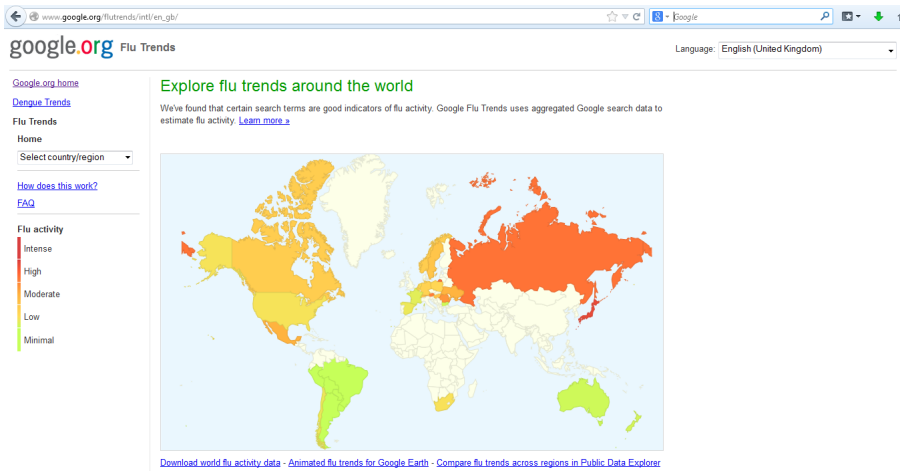

Fig. 4 Google Flu Trends to estimate flu activity during two weeks. Screenshot taken from the Google Flu Trends Web (http://www.google.org/flutrends/), online accessed on 18th March 2014.

System Name	Website	Data Sources	Description
Global Public Health Intelligence Network (GPHIN)		News wires and Web Sites	Warning tool to disease outbreaks
BioCaster	http://born.nii.ac.jp	RSS feeds, Google News, WHO, ProMED-mail and European Media Monitor	Ontology-based system for monitoring online media data
HealthMap	http://www.healthmap.org	Google News, ProMED-mail, WHO and Euro Surveillance	Global disease alert map
EpiSpider	http://www.epispider.org/	ProMed-mail and medical web sites	Human and animal disease alert map
MedISys	http://medusa.jrc.it/medisys/homeedition/en/home.html	Articles concerning public health from news	Monitoring tool for human and animal infectious diseases and chemical, biological, radiological and nuclear threats
Google Flu Trends	http://www.google.org/ flutrends/	Google search and CDC reports	Monitoring system of influenza

Table 1 Epidemic Intelligence Systems.

the real world given a stream of textual or multimedia data from social media. Web mining is the application of data mining techniques to discover and retrieve useful knowledge from the web documents and services. Therefore, the application of these techniques to knowledge extraction provides a better using and understanding of the data space on biomedical and health care domain [29].

There are several health data sources very useful to detect and prevent new outbreaks of different diseases. Social web media and web sites give a large amount of useful data for this purpose. Other important data sources are search engines such as Google and Yahoo! [15] [26]. In this case, the objective is to detect specific searches that involve terms that indicate influenza-like-illness (ILI) through the keywords of the queries performed. The complexity is to interpret the search context of the query, as the user may query about a particular drug, symptom or illness for a variety of reasons. Finally, ProMED-mail [28] is also a widely data source used for disease

outbreaks detection. It is a human network of expert volunteers operating 24/7 as an official program of the International Society for Infectious Diseases. Their volunteers monitor global media reports and in many cases have a reporting time of outbreak disease alerts better than WHO reports.

Text mining techniques have been applied on biomedical text corpus to named entity recognition, text classification, terminology extraction, or relationship extraction [9]. These methods are a human language processing algorithms that aim to convert unstructured textual data from large-scale collections to a specific format filtering them according to the needs.

Once the data have been extracted from the social media sites (RSS feeds, WWW, social networks, ProMED-mail, search engines, etc...), the next stage is to perform the text analysis methods for the trend detection, identifying potential sources of disease outbreaks. These methods can be used to detect words related to diseases or their symptoms in published texts [20]. But this goal can be difficult because the same word can refer to a different thing depending upon context. Furthermore, a specific disease can have multiple names and symptoms associated which increases the complexity of the problem. Ontologies can help to automate human understanding of key concepts and relations between them and allow that a level of filtering accuracy can be achieved. Biomedical ontologies contain lists of terms and their human definitions, which are then given unique identifiers and classified into classes with common properties according to the specific domain treated. In the domain of EI it is necessary to identify and link term classes such as disease, symptom and species in order to detect potential focus diseases. Currently there are various available ontologies that contain all the biomedical terms necessary. For example, BioCaster ontology (BCO) [12] is in the OWL Semantic Web language to support automated reasoning across terms in 12 languages.

A new unsupervised machine learning approach to detect public health events is proposed in Fisichella et al. work [14] which can complement existing systems since it allows to identify public health events (PHE) even if no matching keywords or linguistic patterns can be found. This new approach defines a generative model for predictive event detection from document by modeling the features based on trajectory distributions.

Discovering time and location of the text is the value added by EI systems for high quality. In practice location names are often highly ambiguous because geo-temporal disambiguation is so difficult, and because of the variety of ways in which cases are described across different texts. Keller et al. [19] work provides a review of the issues for epidemic surveillance and present a new method for tackling the identification of a disease outbreak location based on neural networks trained on surface feature patterns in a window around geo-entity expressions.

Finally, a different solution for outbreak detection is shown in Leskovec et al paper [22], where the problem is modelled as a network in order to detect the spreading of the virus or disease as quickly as possible. They present a new methodology for selecting nodes to detect outbreaks of dynamic processes spreading over a graph. This work shows that many objective functions for detecting outbreaks in networks, such as detection time, likelihood, and population affected, are submodular. This

means that, for instance, reading only a few blogs provides more new information than reading it after we have read many ones. They use this characteristic to develop an efficient approximation algorithm (CELF) which achieves near-optimal solutions and it is 700 times faster than a simple greedy algorithm.

4 Twitter applications for tracking and monitoring epidemics

The increasing popularity and use of micro-blogging services such as Twitter are recently a new valuable data source for web-based surveillance because of its message volume and frequency. Twitter users may post about an illness, and their relationships in the network can give us information about which people could be in contact with. Furthermore, user posts retrieved from the public Twitter API can come with GPS-based location tags, which can be used to detect potential disease outbreaks for a health surveillance system.

Recently, several works have already appeared shown the potential of Twitter messages to track and predict disease outbreaks. Ritterman et al. [27] work is focused on using prediction market to model public belief about the possibility that H1N1 virus will become a pandemic. In order to forecast the future prices of the prediction market, they decided to use the Support Vector Machine algorithm to carry out regression. A document classifier to identify relevant messages is presented in Culotta et al. paper [13]. In this work, Twitter messages related to flu were recollected during 10 weeks using keywords such as flu, cough, sore throat or headache. Then, several classification systems based on different regression models to correlate these messages with CDC statistics were compared, finding that the best model achieves a correlation of 0.78 (simple model regression).

Aramaki et al. [1] presents a comparative study of various machine-learning methods to classify tweets related to influenza into two categories: positive or negative. Their experimental results show that SVM model using a polynomial kernel achieves the highest accuracy (FMeasure of 0.756) and the lowest training time.

A novel real-time surveillance system to detect cancer and flu is described in paper [21]. The proposed system continuously extracts text related the two specific diseases from twitter using Twitter streaming API and applies spatial, temporal, and text mining to discover disease-related activities. The output of the three models is summarized as pie charts, time-series graphs, and US disease activity maps on the project website as can be seen in Figure 5. This system can be useful not only for early prediction of disease outbreaks, but also for monitoring distribution of different cancer types and the effectiveness of the treatments used.

Well known regression models are evaluated on their ability to assess disease outbreaks from tweets in Bodnar et al. [4]. Regression methods such as Linear, Multivariable an SVM, are applied to the raw count of tweets that contain at least one of the keywords related to a specific disease, in this case ”flu”. The results confirmed that even using irrelevant tweets and randomly generated datasets, regression methods were able to assess disease levels comparatively well.

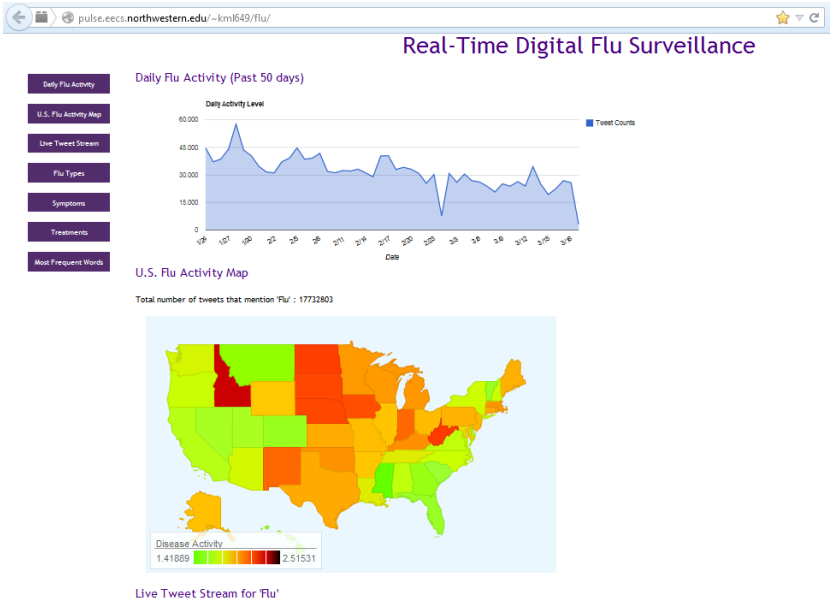

Fig. 5 A novel real-time surveillance system to detect cancer and flu from Twitter messages. Screenshot taken from the Project Web (http://pulse.eecs.northwestern.edu/ kml649/flu/), online accessed on 19th March 2014.

Finally a summary of all the systems mentioned and the machine learning techniques used is listed in Table 2. It can be noticed that most of the works use regression models, and are usually focused on detecting influenza outbreaks.

Work Name	Machine Learning Techniques	Description
Ritterman et al.	Prediction market model and SVM	Predict flu outbreak detection
Culotta et al	Regression models	Classifier to identify flu relevant messages
Aramaki et al.	SVM using a polynomial kernel	Classifier of influenza tweets into positive or negatives
Lee et al.	Spatial, temporal, and text mining	Surveillance system to detect cancer and flu
Bodnar et al.	Regression models	Disease outbreak detection

Table 2 Tracking and monitoring epidemic works using Twitter data.

5 Discussion

All the systems and solutions presented have demonstrated the successful and beneficial use of artificial intelligence techniques when applied to extract and acquire new knowledge for public healthcare purposes. The main challenge of these systems is to interpret the search context of a particular query or document, because an user can query about a particular drug, symptom or illness for a variety of reasons. This goal can be difficult because the same word can refer to a different thing depending upon context. Furthermore, a specific disease can have multiple names and symptoms related to it, which increases the complexity of the problem. Therefore,

to develop strategies for reducing false alarms and decreasing percentage of irrelevant events detected by the epidemic systems can be an important issue for future works and researches on the field.

Additionally, to identify the time and location of messages is a value added for increasing the quality of detecting possible new diseases outbreaks. But in practice location names are often highly ambiguous because geo-temporal disambiguation is so difficult, and because of the variety of ways in which cases are described across different texts.

There are several recent works show the potential of Twitter to track and detect disease outbreaks. These works demonstrate that there are health evidences in social media which can be detected. But, there can be complications regarding the possible incorrect predictions because of the huge amount of social data existing compared with the small amount of relevant data related to potential diseases outbreaks. Therefore, it is necessary to test and validate carefully all the models and methods used.

Acknowledgements This work was supported by Spanish Ministry of Science and Education under Project Code TIN2010-19872 and Savier Project (Airbus Defence & Space, FUAM-076915)

References

1. Aramaki, E., Maskawa, S., Morita, M.: Twitter catches the flu: detecting influenza epidemics using twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1568–1576. Association for Computational Linguistics (2011)
2. Asur, S., Huberman, B.A.: Predicting the future with social media. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, vol. 1, pp. 492–499. IEEE (2010)
3. Bello, G., Menéndez, H., Okazaki, S., Camacho, D.: Extracting collective trends from twitter using social-based data mining. In: Computational Collective Intelligence. Technologies and Applications, pp. 622–630. Springer (2013)
4. Bodnar, T., Salathé, M.: Validating models for disease detection using twitter. In: Proceedings of the 22nd international conference on World Wide Web companion, pp. 699–702. International World Wide Web Conferences Steering Committee (2013)
5. Brownstein, J.S., Freifeld, C.C., Reis, B.Y., Mandl, K.D.: Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project. PLoS medicine **5**(7), e151 (2008)
6. Carneiro, H.A., Mylonakis, E.: Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clinical infectious diseases **49**(10), 1557–1564 (2009)
7. Chen, H., Zeng, D.: Ai for global disease surveillance. Intelligent Systems, IEEE **24**(6), 66–82 (2009)
8. Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1029–1038. ACM (2010)
9. Cohen, A.M., Hersh, W.R.: A survey of current work in biomedical text mining. Briefings in bioinformatics **6**(1), 57–71 (2005)
10. Collier, N.: Uncovering text mining: A survey of current work on web-based epidemic intelligence. Global public health **7**(7), 731–749 (2012)

11. Collier, N., Doan, S., Kawazoe, A., Goodwin, R.M., Conway, M., Tateno, Y., Ngo, Q.H., Dien, D., Kawtrakul, A., Takeuchi, K., et al.: Biocaster: detecting public health rumors with a web-based text mining system. Bioinformatics **24**(24), 2940–2941 (2008)
12. Collier, N., Goodwin, R.M., McCrae, J., Doan, S., Kawazoe, A., Conway, M., Kawtrakul, A., Takeuchi, K., Dien, D.: An ontology-driven system for detecting global health events. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 215–222. Association for Computational Linguistics (2010)
13. Culotta, A.: Towards detecting influenza epidemics by analyzing twitter messages. In: Proceedings of the first workshop on social media analytics, pp. 115–122. ACM (2010)
14. Fisichella, M., Stewart, A., Cuzzocrea, A., Denecke, K.: Detecting health events on the social web to enable epidemic intelligence. In: String Processing and Information Retrieval, pp. 87–103. Springer (2011)
15. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature **457**(7232), 1012–1014 (2009)
16. Hartley, D.M., Nelson, N.P., Walters, R., Arthur, R., Yangarber, R., Madoff, L., Linge, J., Mawudeku, A., Collier, N., Brownstein, J.S., et al.: The landscape of international event-based biosurveillance. Emerging Health Threats **3** (2010)
17. Kamel Boulos, M.N., Sanfilippo, A.P., Corley, C.D., Wheeler, S.: Social web mining and exploitation for serious applications: Technosocial predictive analytics and related technologies for public health, environmental and national security surveillance. Computer Methods and Programs in Biomedicine **100**(1), 16–23 (2010)
18. Keller, M., Blench, M., Tolentino, H., Freifeld, C.C., Mandl, K.D., Mawudeku, A., Eysenbach, G., Brownstein, J.S.: Use of unstructured event-based reports for global infectious disease surveillance. Emerging infectious diseases **15**(5), 689 (2009)
19. Keller, M., Freifeld, C.C., Brownstein, J.S.: Automated vocabulary discovery for geo-parsing online epidemic intelligence. BMC bioinformatics **10**(1), 385 (2009)
20. Lampos, V., Cristianini, N.: Nowcasting events from the social web with statistical learning. ACM Transactions on Intelligent Systems and Technology (TIST) **3**(4), 72 (2012)
21. Lee, K., Agrawal, A., Choudhary, A.: Real-time disease surveillance using twitter data: demonstration on flu and cancer. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1474–1477. ACM (2013)
22. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 420–429. ACM (2007)
23. Linge, J.P., Belyaeva, J., Steinberger, R., Gemo, M., Fuart, F., Al-Khudhairy, D., Bucci, S., Yangarber, R., van der Goot, E.: Medisys: Medical information system. Advanced ICTs for Disaster Management and Threat Detection: Collaborative and Distributed Frameworks pp. 131–142 (2010)
24. Mykhalovskiy, E., Weir, L.: The global public health intelligence network and early warning outbreak detection. Canadian journal of public health **97**(1) (2006)
25. Paquet, C., Coulombier, D., Kaiser, R., Ciotti, M.: Epidemic intelligence: a new framework for strengthening disease surveillance in europe. Euro surveillance: bulletin europeen sur les maladies transmissibles= European communicable disease bulletin **11**(12), 212–214 (2005)
26. Polgreen, P.M., Chen, Y., Pennock, D.M., Nelson, F.D., Weinstein, R.A.: Using internet searches for influenza surveillance. Clinical infectious diseases **47**(11), 1443–1448 (2008)
27. Ritterman, J., Osborne, M., Klein, E.: Using prediction markets and twitter to predict a swine flu pandemic. In: 1st international workshop on mining social media (2009)
28. Victor, L.Y., Madoff, L.C.: Promed-mail: an early warning system for emerging diseases. Clinical infectious diseases **39**(2), 227–232 (2004)
29. Xie, Y., Chen, Z., Cheng, Y., Zhang, K., Agrawal, A., Liao, W.K., Choudhary, A.: Detecting and tracking disease outbreaks by mining social media data. In: Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, pp. 2958–2960. AAAI Press (2013)